Poisoning Whisper: Exploring Targeted Hallucinations in Automatic Speech Recognition

Simon Knight

MInf Project (Part 2) Report
Master of Informatics
School of Informatics
University of Edinburgh

2025

Abstract

This dissertation investigates the phenomenon of hallucinations in large-scale Automatic Speech Recognition (ASR) systems, with a focus on OpenAI's Whisper model. It begins by replicating and extending the findings of Careless Whisper (Koenecke et al.), verifying that hallucinations in Whisper are a recurring and unpredictable phenomenon, and classifying their potential harms. It further identifies novel hallucination behaviours in Whisper-Turbo, including the emergence of fabricated context prompts not previously documented. Building on these findings, the second half of the work explores inducing hallucinations in Whisper. More specifically, whether targeted hallucinations conditioned on speaker accent can be introduced through fine-tuning. Although accent-specific hallucinations were not successfully achieved, the experiments show that hallucinatory behaviour can be modulated to some degree. The work concludes that while practical Hallucination inducing data poisoning attacks on ASR systems remain challenging, they are theoretically plausible and warrant further investigation. This dissertation contributes experimental methods, technical observations, and new questions to the growing field of ASR robustness and safety.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Simon Knight*)

Acknowledgements

I would like to thank my family and my siblings, Jonathan, Robert, Andrew and Eleanor, for being an inspiration to me these past 4 years. They make me who I am more than anything else. I would also like to thank friends in Edinburgh and elsewhere, Your support has been critical. I would also like to thank my ever-suffering Supervisor Peter Bell, who has advised me through the course of my dissertations in spite of my habitual tardiness.

Table of Contents

Chapter 1

Introduction

This research is concerned with the exploration of hallucinations and bias in very large Automatic Speech Recognition (ASR) systems by examining the hallucinatory behaviour of OpenAI's Whisper model, with a particular focus on whether malicious fine tuning, or deliberately introducing biased or misleading patterns during model adaptation can induce hallucinations biased by speaker accent.

1.1 Motivation

Large-scale ASR models like Whisper have been widely praised for their robustness and multilingual capabilities gained from training on hundreds of thousands of hours of diverse web-sourced data.[26] However, this breadth of data introduces risks of erroneous outputs. One particularly troubling failure mode is hallucination. Hallucination is behavior where a model transcribes words or phrases not present in the audio, degrading accuracy. While hallucinations are well-documented in text based large language models, they remain relatively understudied in ASR, despite severe consequences if they occur in ASR deployment in high-stakes domains like journalism or law.

This dissertation builds on two threads of prior work. First, it follows on from the author's undergraduate research, which studied Whisper's transcription quality across English dialects and examined bias demonstrated by Whisper. Second, it takes direct inspiration from Koenecke et al.'s recent study, Careless Whisper[15] (CW), which showed that Whisper not only hallucinates at a non trivial rate (1.1%) but that these hallucinations can produce harmful and misleading content, especially in transcripts of individuals with speech impairments. Koenecke's work highlighted both the scale and the risks of these hallucinations, and raise many questions, which this dissertation aimed to explore.

1.2 Summary of UG1 work

The topic undertaken in the previous year, "*Exploring the bias in very large automatic speech recognition systems*"[13], investigated bias on OpenAI's Whisper model. The

work explored whether Whisper performs differently based on speakers' linguistic backgrounds and accents and through this analysis highlighted systemic bias that manifested in transcription. Two key datasets were used for testing: the Edinburgh International Accents of English Corpus (EdACC)[3] and the George Mason University Speech Accent Archive (SAA)[33]. These were chosen for their high audio quality and rich metadata, as well as their diversity in speaker accents and first languages.

The study conducted a series of experiments across various speaker groups from the EdACC and GMU corpora, using a range of sizes of Whisper models (Tiny to Medium), evaluating performance degradation under noise, and examining the model's built-in language detection features. Whisper's transcription outputs were compared against gold-standard references and normalized. Bias was examined at multiple levels, from individual languages to linguistic families and genera, using groupings sourced from the WALS database[8].

Results revealed that Whisper had unusual performance differences in performance between different groups of English 2nd language speakers from a wide range of linguistic backgrounds. Notably, speakers from Romance language backgrounds and East Asian languages exhibited higher Word Error Rates (WER). These differences in WER between groups were non-trivial. Noise-level testing confirmed Whisper's resilience to audio interference but also further demonstrated performance disparities between speakers in noisy conditions.

Ultimately, the work highlights meaningful patterns of bias in Whisper while acknowledging limitations related to dataset size, resource constraints, and model selection. The findings suggest that even advanced models like Whisper may require further development to ensure equitable performance across diverse user groups, and offered a robust experimental framework for further study.

As it was completed as an UG1 project, the work completed was not initially developed with the understanding that it would be expanded upon in further work. However, The work completed this year expands upon analysis within the realm of bias in ASR systems, with specific expansion of hallucinations, which were observed but remained unexamined in last year's work.

1.3 Problem Statement and Contribution

This dissertation is intended as an investigation of hallucinatory behaviours demonstrated by Whisper, starting initially with building an understanding of hallucination as it manifests itself in the largest and most recent models created by OpenAI, and moving onto experimentation into ways in which Whisper could hallucinate a targeted phrase in response to speech from speakers with particular speech characteristics. We believed this would act as a proof-of concept for a poisoning attack which could be scaled up to affect the training of Whisper itself.

In doing so, this work contributes the following

- A replication and validation of Koenecke et al.'s findings using a separate corpus

(EdAcc)[3], including hallucination rate analysis, statistical modelling, classification of hallucination harms, and confirming speaker harms.

- An analysis of novel hallucinatory behaviour demonstrated by Whisper Turbo, not been reported on in prior literature.
- A targeted exploration of speaker-biased hallucination in Whisper through fine-tuning, yielding inconclusive results and highlighting the difficulty of reliably inducing such behaviour.
- An experimental framework for training Whisper to hallucinate trigger phrases based on speaker characteristics.
- A technical failure analysis identifying where the experimental design did not produce the expected results, with discussion of potential causes and refinements.

Ultimately while successful in the replication and expansion of ideas introduced in Careless Whisper[15], the extent of the discoveries in this paper has been affected by issues related to resource limits and suboptimal decision making in early stages of experimentation. It is hoped that the experiments can be expanded on in the future, with pitfalls experienced in our experimental process avoided.

1.4 Dissertation Structure

This dissertation is organised into the following Chapters:

Chapter 2 introduces Whisper's architecture, training data composition, and its known limitations. Reviews literature on hallucinations in ASR and data poisoning attacks in NLP and speech systems.

Chapter 3 describes datasets used, including CommonVoice[20], EdAcc[3], VCTK[30], and those excluded. Discusses practical challenges such as accent annotation, recording quality, and speaker metadata availability.

Chapter 4 details the replication of Koenecke et al.'s[15] hallucination study on the EdAcc corpus using Whisper Large-V3 and Turbo, including harm classification and foreign language hallucination behaviors.

Chapter 5 outlines implementation and development of concepts, datasets and models utilised in the Results and analysis of poisoning experiments, such as model fine-tuning configurations, noise-controlled variants, and the structure of test datasets. Also reflects on key methodological limitations.

Chapter 6 presents results from poisoned Whisper models. Examines hallucination rates across speaker groups and test conditions, reflects on technical errors, and proposes reasons for inconsistent outcomes.

Chapter 7 briefly summarises the main conclusions and findings of this paper.

Chapter 2

Background

This Chapter is intended as a summary of key concepts and methodologies necessary to understand the work presented in this dissertation. It also provides an overview of relevant prior research and related developments in the field, highlighting contributions that have influenced the direction or progression of this study.

2.1 Whisper

This section of the Background Covers Whisper[26], OpenAI's cutting edge Automatic Speech Recognition (ASR) system.

Whisper, developed by OpenAI, is designed to transcribe spoken word into text with high precision. It achieves this through training on an exceptionally large and diverse multilingual dataset that represents the largest amount of training data yet used on an ASR model, totalling more than 600,000 hours of transcribed data. Rather than relying solely on curated gold-standard datasets, OpenAI sourced most of this data from platforms such as YouTube. OpenAI claims this trade-off has enabled Whisper to develop a model with significantly enhanced robustness and generalisation capabilities when compared to other commercial systems. For example, it demonstrates strong performance on both transcription and translation tasks, even in unfamiliar or challenging contexts[27], contributing to its widespread adoption. Whisper is not without its drawbacks however, as its reliance on web data introduces potential biases and transcription inconsistencies, particularly for low resource languages, or less common accents. This has been seen in work by Graham et al.[12], as well as our own work completed last year, *"Exploring the Bias in Very Large Automatic Speech Recognition Systems"*.

A more general overview of Whisper was provided in this previous paper. To avoid retreading ground, this background section will focus only on concepts not previously introduced, as well as those necessary for understanding the specific contributions and focus of this dissertation.

2.1.1 Whisper's architecture

Whisper is built on a sequence-to-sequence neural network architecture that employs transformers, structured as an attention-based encoder-decoder model with self-attention mechanisms. The system first processes input audio through convolution layers for initial feature extraction. The extracted features are then processed through a series of transformer blocks that use self-attention to identify and prioritise the most relevant elements for transcription, enabling the model to learn contextual dependencies. The decoders then use cross-attention to translate the encoded representations into text. This unified architecture allows Whisper to perform effectively at both transcription and translation, and also improves its overall generalization and efficiency.

The attention-based encoder-decoder model forms the backbone of many recent high performing speech recognition systems including Whisper, but also models such as Nvidia's Canary[16] and Meta's Seamless[9]. These Architectures are favoured over older ASR approaches that rely on multi-model pipelines, such as the DNN-HMM hybrid design, which separate the acoustic modelling and language modelling into distinct components. These older systems require that each of these components are trained independently, and were favoured as they required a lower amount of data to train. With the advent of transformer based models and the growing availability of resources, however, we have seen a shift towards unified architecture. These models offer improved generalization across tasks and accents, even on unseen data, while also simplifying fine-tuning for specific functions. However, this transition has introduced certain challenges: The data and compute requirements to create these models have increased exponentially, becoming difficult to reproduce without the resources available to large organisations. Some of the broader implications of this are explored further in the section on data poisoning.

2.1.2 Whisper since UG1

Since December 2023, OpenAI has released several updates to Whisper, most notably the introduction of Whisper Turbo[24], the new flagship model for Whisper. This optimised version is designed to deliver significantly faster inference speeds all the while at the cost of lower computation, giving greater efficiency, thereby bridging the gap between the high-performance of the large models and the accessibility of the smaller variants. These improvements are primarily achieved by reducing the number of decoder layers from 32 to just 4, cutting the models parameter count by nearly half, while keeping the encoder architecture the same. As a result, inference speeds have increased five times. The obvious downside of this is that Whisper-Turbo exhibits a slight reduction in accuracy compared to the Largev3 model.

2.2 Data Poisoning

Data poisoning is a form of adversarial attack in which malicious actors deliberately introduce misleading, biased or incorrect data into a training dataset in order to manipulate the behaviour of machine learning models that are trained on that data. It is

distinguished from harmful but unintentional behaviour by the fact that the degradation of the model performance is a direct result of a targeted interference.

One of the first papers to explore this was Biggio et al. [4], who introduced a foundational analysis of poisoning attacks against Support Vector Machines. Their work demonstrated how carefully crafted training samples could significantly degrade model accuracy, establishing the broader theoretical underpinnings for adversarial data manipulation in machine learning.

In contrast to this theoretical groundwork, one of the earliest and most widely documented cases of real world data poisoning in language models occurred in 2016 with Microsoft's Tay chatbot[31]. Tay was a conversational agent deployed on Twitter (now X) that interacted with users and adapted its language patterns and responses via a self-learning loop. This design, while innovative, left the system vulnerable to adversarial manipulation. Twitter users proceeded to intentionally expose Tay to inflammatory and inappropriate content in a coordinated manner [19], which the model subsequently internalized and echoed. The lack of content filtering and safeguards allowed the attacks to influence outputs significantly, leaving Microsoft with little option but to withdraw the system. The incident underscored the dangers of training models in the wild without suitable control and defences of the input.

Since the Tay incident, research into data poisoning in language models has expanded ins several distinct directions. One avenue of study investigates stealthy and hard to detect poisoning techniques. For example, a paper in 2021 by Shumailov et al.[28] successfully demonstrated that model behaviour could be subtly biased merely by altering the order of training samples, reducing required modification for poisoning, and demonstrating the increasing subtlety of attacks. Another branch of research explores trigger based attacks. Wallace et al.[32] showed that language models could be conditioned to respond in specific ways when exposed to predefined trigger phrases, that when embedded into input queries, could reliably trigger biased or incorrect responses, showcasing ways that the adversarial behaviour could present itself.

The mechanisms through which data poisoning attacks are executed are comprehensively outlined in Steinhardt et al.[29] , a paper that offers a foundational analysis of poisoning strategies and their impact on model robustness. The study categorises various poisoning methods and evaluates their practical effects on different model types.

The increasing Mitigation strategies for data poisoning include the implementation of stricter data validation procedures, the use of adversarial training techniques and the adoption of more rigorous standards in dataset curation to ensure the integrity of the training data.

2.2.1 Data Poisoning in ASR systems

Adversarial attacks against ASR systems have historically focused on evasion attacks, where maliciously crafted audio samples cause misclassification at inference time, through techniques that involve intercepting input data and modifying it. However, as ASR systems increasingly use web-sourced or loosely curated training datasets, there has been growing interest in poisoning attacks that occur during the training phase.

Training-time data poisoning presents particular challenges for detection and mitigation, as the manipulated data can be embedded deeply into the model's learned behaviour often times being indistinguishable from poisoned data

While data poisoning is well documented in text-based natural language processing(NLP) models, its impact on ASR systems remains a relatively under explored area. The number of published studies on ASR poisoning is considerably smaller than in adjacent fields such as text large language models (LLMs). We hypothesise that this is because of a relative lack of novelty and interest surrounding ASR. Unlike traditional language models that rely on textual input, ASR models must process acoustic signals and their transcriptions, which makes them susceptible to new attack vectors.

One prominent form of poisoning in ASR models involves manipulating the transcription data used during training. Kurakin et al. (2018) were able to demonstrate that introducing incorrect transcriptions into large-scale training datasets can systematically degrade the accuracy of ASR models[17]. The researchers were able to find that over time, the poisoning reduced transcription reliability and increased error rates, particularly in challenging audio conditions. They found that this strategy was particularly effective in the context of web-scraped corpora, where insufficient data validation practices allowed poisoned transcripts to be incorporated unnoticed into training pipelines, which had devastating results in the behaviour of the model.

An alternative poisoning strategy involves the manipulation of the audio signal itself rather than its transcription. For example, Cai et al. proposed a method using high-frequency trigger signals to subtly modify input audio in ways imperceptible to human listeners or basic automated filters[7]. However, more interesting to the researchers of this paper was the use of Voice Style Transfer techniques, in which researchers used a model to extract timbral-related characteristics from one speaker, and used a voice conversion model to modify a speaker to make fluent and natural audio with only timbral differences. While for their purposes, they primarily observed models failing at classification challenges, this reveals that behaviours of ASR models can change with a difference of certain speaker traits.

This is also seen in Yao et al.'s work[35], where the authors subtly modify the rhythm of speech in selected audio files while preserving other qualities such as timbre and linguistic content. Their results show a high attack success rate, with models consistently misclassifying these adversarial samples. As speech timbre and rhythm are key components of regional elements of speech, we believe that these papers together present a credible foundation for the hypothesis that ASR behaviour can be influenced or triggered by the accent of the speakers.

A notable study in ASR is Venomave, an adversarial poisoning attack targeting ASR systems through vulnerabilities in dataset curation[1]. Venomave differs from earlier techniques by systematically manipulating both the acoustic and textual components of the training data in ways that are difficult to detect. This dual-modality manipulation allows it to bypass common data validation methods, making it a particularly potent and stealthy form of attack.Venomave exploits the nature of hybrid ASR architectures, targeting the Acoustic DNN by introducing imperceptible perturbations into the training audio, which bias the acoustic deep neural network (DNN) during

phoneme recognition[1]. These corrupted phoneme predictions then propagate through the system, ultimately disrupting the phoneme-to-word mappings handled by the HMM-based language model. Alarmingly, the paper's authors were able to retain an 80% success rate of attack when the percentage of poisoned training samples was reduced down to just 0.17% of the samples. Most interestingly, this poisoned data was found to be remarkably transferable to models of different structure, with a unified end-to-end model found to exhibit target behavior in 36.4% of transcripts.

2.3 Exploring Hallucinations in ASR

This section examines *Careless Whisper: Speech-to-Text Hallucination Harms* by Koenecke et al.[15], a paper that directly informed the direction of this dissertation and built upon questions explored in our previous undergraduate research. This paper was of particular interest to us given that Koenecke, its lead author, is a leading figure in the study of bias in ASR and broader AI systems. Moreover, the themes explored in Careless Whisper closely align with those of our previous dissertation. While answering many questions raised in our previous work, Koenecke's work also opened up potential paths for exploration that directly informed the direction of this dissertation.

The paper by Koenecke et al. explores the phenomenon of hallucinations in OpenAI's Whisper, drawing attention to the real-world risks these errors pose. In the context of generative AI, a hallucination refers to the unintended generation of incorrect or fabricated information that is not present in the input data, a phenomenon more commonly observed and studied in large language models like ChatGPT[2]. In speech-to-text systems, hallucinations occur when the model transcribes words or sentences that were never spoken in the original audio. Unlike transcription errors, such as mis-recognising a word due to poor audio quality, hallucinations involve the fabrication of entirely new content.

Although hallucinations have been widely recognised in text LLMs, where factual inaccuracies are easily identified, they remain relatively underexplored in the context of ASR systems. This gap in research seems to persist in spite of the obvious danger of ASR hallucination, such as the attribution of incorrect speech to an individual, which would be particularly dangerous in contexts such as legal proceedings or journalism. The difficulty in identifying hallucinations stems from the challenge of distinguishing them from standard transcription errors. This was explored by Frieske et al.[10] in their work on ASR hallucination, which defines these errors as outputs that are semantically unrelated to the input and yet are linguistically fluent. They argue that Word Error Rate (WER), the most used evaluation metric for ASR systems, is not as useful for detecting hallucinations. They Identify that hallucinatory models can be identified by observing how WER changes when controlled perturbations are introduced to the input audio: hallucination-prone models tend to show greater WER fluctuations under such conditions. Despite this, identifying hallucinated content in practice still requires manual transcript-by-transcript review, as there is no automated metric currently for identifying veracity.

returning to Koenecke et al, they found that Whisper produced entirely novel phrases

and sentences in approximately 1.2% of transcripts. Despite Whisper's lower WER compared to other ASR systems on standardized tests, the model was uniquely prone to hallucinations on benchmark datasets that did not affect competitor models such as Google's Chirp. Koenecke goes on to identify that the rate of the hallucinations are higher in individuals with aphasia, a speech disorder that is characterized by a difficulty in producing spoken language.

Their large-scale evaluation used over 13,000 audio segments from AphasiaBank, a repository of speech recordings from individuals with aphasia, as well neurotypical control speakers. The authors used more than 13,000 audio segments in their work and found that in contrast to OpenAI's public claims about the accuracy and reliability of Whisper on mainstream speech benchmarks, hallucinations occurred in 1.2% of all transcripts, and were able to further identify that roughly 40% of these hallucinations exhibited behaviour that the authors classified as harmful.

For classification of harms, 3 overarching categories of harm were described to aid identification. The first category, inaccurate associations, involves Whisper incorrectly referencing names, locations, or medical conditions, potentially leading to negative connotations or the spread of misinformation. The second category, perpetuation of violence, includes hallucinations containing aggressive or inappropriate content such as references to violence, threats, sexual language, or criminal acts. The third, false authority, refers to hallucinations that mimic authoritative statements, including fabricated citations, names of non-existent organizations, or false attributions. A notable subcategory of this was described as video-based authority, where hallucinated content mimicked language commonly found in video media. for example, phrases like "Thanks for watching" or "You guys at home, you know exactly what we're about to say."

Harm Category	**Representative Hallucinated Example**
Inaccurate Associations	"Mike was the PI, Coleman the PA, and the leader of the related units were my uncle..."
Perpetuation of Violence	"...I'm sure he didn't have a terror knife so he killed a number of people..."
False Authority	"To learn more, please visit SnowBibbleDog.com."

Table 2.1: Examples of hallucination harms by category from Whisper transcriptions

The authors emphasised the serious implications of such errors, particularly in high stake contexts such as medicine and law, where even minor transcription inaccuracies can have major consequences. Importantly, the hallucinations were found to be non-deterministic, and non-replicable across repeated runs, highlighting the unpredictability of Whisper's behaviour. Additionally they also could not identify any obvious links between the speaker characteristics and any generated hallucinations or the types of harm that were being perpetuated in their text, barring a slight increase of hallucination rate in those with aphasia.

The authors propose 2 hypotheses for why Whisper exhibits this behaviour. They suggest that Whisper, being a seq2seq model created by openAI, may share some underlying architectural features with other OpenAI models, such as the hallucination

prone ChatGPT. Second, they suggested that the nature of Whisper's training data may contribute to the appearance of "false authority" hallucinations. This is supported by the fact that many hallucinated outputs resembled scripted language similar to content from video subtitles, online lectures, or similar sources that may have been overrepresented in the training corpus.

An important finding that the authors et al also establish is conditions in audio that increase the likelihood of a hallucination. They found that recordings of individuals with speech impairments, such as aphasia were disproportionately affected. Speakers with longer non-vocal durations (e.g., pauses or disfluencies) were more likely to experience hallucinations in their transcriptions As the hallucinations manifested themselves primarily in speech with gaps of silence (with ambient noise). Koenecke et al. ends with a call to action to OpenAI to reduce the harms that can be caused by hallucinations output by Whisper, with a focus on people with aphasia.

2.4 From Bias to Hallucinations: An Emerging Research Direction

In our UG4 project, we set out to investigate bias in Whisper ASR. We had findings related to bias in Whisper that we found convincing. Still, scattered throughout our dataset, we encountered transcripts that clearly bore no relation to the original audio, which we recognised as hallucinations. These initially appeared to offer a potential vector through which Whisper could demonstrate biased behaviour toward speakers with different dialects or accents. We attempted a preliminary analysis, tracking hallucination frequency across speaker groups and examining content for signs of harmful bias. However, these efforts were ultimately inconclusive, and hallucinations were set aside from the final scope of the dissertation, as the number of identified hallucinations were too small for meaningful statistical analysis, and locating them proved increasingly time-consuming.

It is therefore of particular interest that Koenecke et al. were able not only to identify hallucinations as a recurring phenomenon in Whisper but also to highlight their potential harms and identify broader patterns that manifested through the samples with hallucination in Whisper. In their paper, they lay out new strategies to more efficiently identify hallucinations, and consequently reveal that they are a common phenomenon in Whisper. With this new information, we felt that there was potentially new avenues of exploration for bias in ASR systems.

Although we felt that we needed to replicate the results in order to proceed, we began to ask ourselves questions regarding the content of the paper. According to Koenecke et al. Hallucinations are not biased around speaker traits. What would have to change for this to not be the case? This led us onto exploring the feasibility of training Whisper on poisoned data specifically crafted to induce hallucinations that correlate with speaker characteristics, such as accent or region, and if this suggested the feasibility of a larger attack on Whisper.

Chapter 3

Data

This chapter outlines all datasets used throughout the course of this dissertation, detailing where they were applied within the work, along with implementation specifics where relevant.

We hoped to identify a corpus of spoken English composed of a unified accent of speech. As the author of this paper is from the South East of England accent associated with this region were settled on as this would avoid problem related to misuse and classification of minority accent groups and cultural issues.

The British National Corpus[6] (1994) and MARSEC were initially considered to construct a dataset of received pronunciation (RP) speakers with reliable transcriptions. While they offered a targeted accent group ideal for poisoning experiments, both suffered from poor audio quality and degradation due to outdated recording methods. This raised concerns about the model overfitting to recording artifacts rather than speaker traits, leading to their exclusion from the final dataset.

3.1 CommonVoice

CommonVoice[20] is a large-scale multilingual speech dataset developed by Mozilla, consisting of audio recordings from volunteers worldwide. All 2000+ hours of recordings are short (3–10 seconds) and paired with manually verified transcripts, with a focus on accent and language diversity. The English corpus is especially varied, making it useful for testing model generalisation. CommonVoice 9 was also used in Whisper's original evaluation.

Commonvoice originally formed the core dataset around which the experimental section of this dissertation was conducted, and provided all training, validation, and test material for Whisper fine-tuning, but now serves a reduced role due to apparent issues following closer analysis.

We began by identifying speakers from Southern English regions, using Mozilla's inclusion of accent metadata, starting in CommonVoice 5. The process of volunteer submission of accent data initially involved selecting from a predefined list, but was

later expanded to allow free-text input. Sorting the data revealed over 710 unique accent entries. These were manually filtered to identify 35 accent labels aligned with our criteria, yielding around 202 hours of Southern English/RP recordings. Ultimately, this gave us 202 hours of Southern English/RP accents to train the corpora on.

However, issues soon arose. The accent tags were unstandardised and inconsistent, including entries like “Savage Texas Gentleman,” “blurpy,” and “personal idiolect.” These often reflected idiosyncratic self-identifications and contributed only a few minutes each. Among the 202 hours, 200 came from a single broad category: “England English.” We had hoped to avoid this tag due to its broadness and difficulty to justify as a common accent. This issue is primarily caused by CommonVoice making the option to apply more succinct accent descriptions more involved than selecting from a list of categories.

Next, we manually reviewed 10 minutes of “England English” data and found little to no accent consistency. We believe that many contributors may have misunderstood the label, as this inconsistency was not nearly as present in other predefined categories such as ”Irish English” and ”Singaporean English”. It is likely some users selected “England English” to indicate they were speaking English.

Despite these limitations, the remainder of the CommonVoice 20 corpus was still used in the datasets as the ‘clean’ component of our training datasets. This is because we believed it still held value as a diverse and relatively high-quality source of speech data, particularly for standard English transcription.

3.2 VCTK

The Voice Cloning Toolkit[30] or VCTK is a publicly available speech dataset developed at the University of Edinburgh containing recordings from native English speakers. It includes speech audio from 111 English speaking volunteers, primarily from America or Commonwealth nations. Each speaker reads a script consisting of 400 sentences derived from newspaper articles specifically chosen to maximise the phonetic and contextual coverage, as well as an elicitation paragraph similar to that read by the GMU corpus used last year. The quality of the recordings is high and standardised between speakers, with 2 simultaneous recordings for each speaker elicitation. As duplicate audio would not be useful in our project, only audio from microphone one was used.

Although minimal preprocessing was required overall, one adjustment was made to ensure consistency in file format between VCTK and the CommonVoice corpus. All VCTK audio files were converted from their original .flac format to .mp3 and downsampled to 16 kHz to match the standard audio resolution used throughout the project. This compression was deemed acceptable for the transcription-focused nature of the task.

The VCTK was primarily used throughout this project as a source of reliably identified speakers of Southern/Southern English British. This was possible thanks to the relatively high specificity in the metadata provided by participants regarding their accent, which is information we find is often lacking or inconsistently reported in other public corpora.

The categories used for this subset of the speakers and their number were

- Southern England $\times$ 8
- Surrey $\times$ 2
- South East (SE) England
- South West (SW) England
- Suffolk
- Oxford
- London
- Essex

The use of the VCTK may have caused more issues for us. For one, the speaker spread used above is more broad than would be ideal, which may have affected our success. The relatively small volume of audio may have locked the direction of the experiments, as well as directly influencing a variety of decisions that may have hurt results. This is expanded upon in the methodology section.

3.3 Edinburgh International Accents of English Corpus

The Edinburgh International Accents of English Corpus (EdACC)[3] is a dataset consisting of over 40 hours of speech, with 121 speakers of a variety of linguistic backgrounds speaking conversationally, aiming to provide a more representative sample of diverse speech than previous corpora. The strengths of this dataset are numerous. For one, by using speech recordings of casual conversations, the dataset includes dialect features of speech that are only seen in real conversation and missed in more rigidly formatted datasets. The dataset has been specifically curated to solve the problem of lack of variety in current ASR datasets. The dataset also includes a table of information about the speakers, such as linguistic information, but also includes factors about their background, such as accent self-identification. For the purposes of this dissertation, this was its critical feature.

The transcription of the EdAcc dataset was completed by professional transcribers. This means that the transcription data is of a high standard: every turn in the conversation was manually timestamped and segmented, with multiple non-speech features labelled in the transcriptions, such as speaker overlaps, laughter and others. The high quality of the transcription allows fast and accurate analysis when comparing against Whisper generated transcripts.

This dataset was not used in the fine-tuning experiment section, but was used extensively in testing the hallucinatory capacity of Whisper in Chapter 4.

Chapter 4

Verification and Analysis of Hallucinatory Behaviour in Whisper

This section aims to lay the groundwork for the fine-tuning chapter of this dissertation by verifying the findings of Koenecke et al. [15],identifying trends in hallucinations and also expanding the work to other state of the art Whisper models.

As discussed in the background chapter, Koenecke et al. conducted a landmark investigation into hallucinations generated by Whisper, expanding the study of ASR bias into areas our own analysis had not addressed. Because their work was published while our undergraduate dissertation was being undertaken, we were unable to incorporate their findings into our initial analysis of hallucinations. In their paper, Koenecke et al. reported hallucinations in Whisper transcripts occurring at a rate an order of magnitude higher than anything we observed, and attempted to identify correlations between hallucinations and speaker characteristics. However, they ultimately concluded that hallucination occurrence in Whisper appeared largely random and did not convincingly correlate with speaker traits with the exception of Aphasia.

For the purposes of this dissertation, we felt it was important to replicate Koenecke's experiments, both to verify the extent of Whisper's hallucinatory tendencies under controlled conditions and to establish a reliable baseline from which new hypotheses could be constructed. This also provided an opportunity to expand their findings into newer model architectures released by OpenAI.

A secondary motivation for replication was more personal, rooted in the results of our work from the previous year. In that study, we found that under extreme conditions such as added noise, Whisper sometimes performed worse on specific speaker subsets. Koenecke's reported hallucination rate of 1.1% came as a surprise, as our own work showed only isolated examples. We wanted to revisit this discrepancy by testing hallucination rates on the same dataset used in the previous year's work, the EdACC corpus [3], and assess Whisper's behaviour in direct conversation with our prior findings.

4.1 Methodology

The exact setup of the experiments conducted by Koenecke et al. is not fully detailed in their paper. While they reference the use of the Whisper API, no specifics are given regarding the model size used to generate their results. Given that the paper predates the release of Whisper-Turbo, we presume that the model in question was either Whisper Large-V2 [23] or Large-V3 [22], depending on whether they explicitly used "Whisper-1", or whether the default model in the API had already shifted to Large-V3 at the time of testing. To expand on their findings and examine hallucination tendencies in more recent model variants, we extend our analysis to include Whisper-Turbo and highlight any notable differences in its performance.

4.1.1 Data

A common first step in replicating and validating the results of another study is securing access to the original dataset. The dataset used in Koenecke et al.'s paper is AphasiaBank[18] which provides annotated speech data from individuals with aphasia. However, full access to the complete and unrestricted dataset is limited to members of the AphasiaBank consortium and is password protected. While access could have been arranged through consultation with a university faculty advisor, we determined that a comparable dataset with equivalent speaker variables would allow for a meaningful replication of the study's statistical observations, particularly when tested on a different ASR model.

For this investigation, the EdAcc corpus [3] was selected due to the exceptionally rich speaker metadata it provides, which includes linguistic background, demographic information, and other relevant traits critical to our analysis. This richness of annotation made it particularly well-suited to replicating and expanding the analyses presented in Koenecke et al.'s work. On top of this, as a dataset used in our UG project, we have a baseline familiarity with the dataset.

That said, EdAcc comes with a notable limitation: a smaller speaker pool, comprising 122 unique participants compared to AphasiaBank's 437. This reduced sample size may constrain the statistical significance of our findings, particularly in identifying more nuanced subgroup correlations in hallucination rates. Nonetheless, the depth and precision of EdAcc's speaker metadata justify its selection as the most appropriate available resource for this study.

Table 4.1 shows variables used as independent predictors in Koenecke et al.'s logistic regression analysis. We attempted to identify these variables within the EdAcc corpus, but several caveats apply to our analysis, detailed below:

Vision loss: The EdAcc dataset contains no records related to participants' vision. Koenecke et al.'s results support the idea that Vision has no statistically significant impact on hallucination rates in Whisper. Consequently, this variable was excluded from our analysis.

African American: Koenecke has previously investigated racial disparities in ASR, notably identifying in her 2020 study that commercial systems often performed nearly

twice as poorly for African American speakers compared to white speakers [14]. In this context, it is understandable that their current research continues to attend to this vector of potential harm. However, the EdAcc corpusf does not include any African American participants. The dataset does include racial categories such as White, South Asian, Asian, and Black. However, as African American is both a cultural and racial identity,and the EdACC lacks the detail to capture that nuance, we did not believe that we would be able to draw any useful comparisons with any racial analysis in our work.

Has Aphasia: Although the EdAcc metadata includes indicators for certain speech and hearing impairments, no participant is reported to have aphasia. Furthermore, while hearing loss is tracked, only a single speaker identifies as hearing-impaired. This limited representation made it infeasible to conduct meaningful statistical analysis on either variable, and both were excluded from further consideration.

File specific data: Since file-level features such as silence duration or transcript length are not natively included in the EdAcc metadata, we extracted them manually. Silence duration (in seconds) was computed using PyAnnote [5], while word counts were derived directly from the ground truth transcripts.

Table 4.1: Logistic Regression Predicting Whisper Hallucination, table from Koenecke et al.

	Dependent variable:
	Hallucination
Share of Duration Being Non-Vocal	0.951^{**} (0.438)
Number of Words	0.056^{***} (0.011)
Has Aphasia	0.368^{*} (0.204)
Is Female	-0.017 (0.168)
Age	0.044 (0.043)
Age Squared	-0.0004 (0.0004)
African American	-0.132 (0.469)
Other Race	-0.281 (0.518)
Years of Education	-0.058^{*} (0.033)
English is First Language	-0.227 (1.035)
No Vision Loss	0.395 (1.026)
No Hearing Loss	0.500 (1.028)
Constant	-6.447^{***} (2.175)
Observations	10,830
Log Likelihood	-783.942
Akaike Inf. Crit.	1,593.883
Note:	*p<0.1; **p<0.05; ***p<0.01

4.1.2 Identifying Hallucination

Our initial setup was to verify the rates of hallucination observed in Koenecke et al. on a known dataset.

While the files themselves were unmodified, it was necessary to remove from the test

set all files with a length greater than 30 seconds. This is because Whisper's context window is limited to 30 seconds, as it can only transcribe audio within that span in a single pass, and passing whipser longer clips forces it to internally segment the audio using heuristics that may introduce inconsistencies. Using shorter clips therefore eliminates any potential confounding factors. Of the original 9,848 segments, 6,289 unique segments were used in this analysis of Whisper.

Koenecke et al. noted in their work that the hallucinations generated by Whisper appeared to be non-deterministic. They used this to their advantage along with another separate observation: clips prone to hallucination would often trigger different hallucinated outputs across successive transcriptions.. As hallucinations are non-deterministic, Whisper will generate syntactically different hallucinations with each pass. This enables us to use rules based filtering of transcripts to reduce the work required to identify hallucinations in transcripts. We transcribe all audio twice with each model, using rules similar to those applied by Koenecke et al. to identify hallucinatory transcripts. The rules below were applied disjunctively, with files manually reviewed if they satisfied any condition.

- **Disagreement of 20% in WER or higher between transcripts:** If both transcripts yield hallucinations, their content will be unrelated to each other or to the ground truth transcript, resulting in high disagreement in word error rate (WER).
- **At least 5 insertions from the ground truth in either transcript.**
- **Insertion represents at least 0.5 the length of the original ground truth.**

The second and third rules were used to increase the hallucination candidate catchment in cases where the primary rule failed to identify them. As all files were manually reviewed for false positives, it was rational to overestimate rather than underestimate the number of hallucinations. Note that if a row produced a hallucination, even in only one of the runs, it was included in our hallucination rate over the total number of files.

The manual review process however, is not without flaws. Manual identification of hallucinations can be particularly difficult near the boundaries between transcription errors. We have erred on the side of caution and have been generous with hallucination definitions in our results The results below were gathered by running Whisper LargeV3 and Whisper Turbo, both in transcription mode, with input language set to English.

4.2 Results and analysis

4.2.1 Whisper-LargeV3

With a final hallucination rate of approximately **1.05%** on Whisper Large, our result closely mirrors Koenecke's reported value of **1.1%** on their neurotypical control group. If we assume that Whisper-LargeV2 was used in Careless Whisper [15], then the marginal reduction in hallucinations may suggest that improvements to Whisper's dataset in LargeV3 [22] have helped reduce such behaviour. We think that this difference could also be explained to be subjective difference in hallucination labelling. As

Table 4.2: Logistic Regression Predicting Whisper Hallucination (LargeV3)

Variable	Coefficient	Std. Error	p-value	Significant
Share Non-Vocal (%)	0.104	0.065	0.112	
Number of Words	-0.150	0.055	0.010	**
Is Female	-0.307	0.255	0.229	
Age	-0.002	0.018	0.921	
Years of Education	0.022	0.072	0.756	
English is First Language	0.347	0.276	0.209	
Constant	-2.704	1.202	0.024	**
Observations	6,289	Hallucinations	65	1.05%
Note:				$^{**}p<0.05$; $^{***}p<0.01$

demonstrated in Table 4.3, manually identifying hallucinations posed several challenges. Consequently do not consider the difference in rates to be conclusive of the model used.

Nonetheless, we show that Koenecke's findings can be approximated on a different dataset, reaffirming her claim that hallucinations are a surprisingly common phenomenon in Whisper. A key point of divergence lies in the distribution of hallucination types. We found that only **32%** of hallucinations were clearly harmful. Among these, **24%** were categorised as Perpetuation of Violence, **38%** as False Authority, and **38%** as Inaccurate Association. In contrast, Koenecke observed Perpetuation of Violence as the most prevalent, followed by Inaccurate Association and then False Authority. We could not attribute this shift to any specific feature of the EdAcc dataset.

Our results also broadly align with Koenecke et al's conclusion that speaker traits do not significantly correlate with hallucination rates. However, we observed the opposite of their finding regarding transcript length: while Koenecke noted a positive correlation between the number of words and hallucination , our analysis suggests a negative one. This may be due to our hallucination candidate rule set being more permissive of short transcripts, leading to overrepresentation of short transcripts in our results.

4.2.2 Hallucination Analysis on Whisper-Turbo

Although there is not a point of direct comparison between the work in 'Careless Whisper' and our tests on Turbo, we think that it is a logical expansion of our tests on Whisper, as it expands the tests to the latest developments in ASR technology.

In spite of OpenAI's claims that Whisper-Turbo maintains transcription quality comparable to Large-V3, we observe a hallucination rate of **4.3%** for Turbo, significantly higher than that of Large-V3. We think that alongside a general increase in hallucination due to the smaller model size, a new behaviour is seen in Whisper Turbo which dramatically increases the number of observed hallucinations.

Previously observed hallucination in other models of Whisper primarily had hallucinations of an unpredictable nature. The content of the hallucinations was hard to determine prior to transcription, even on a hallucination-inducing file, with hallucinations on different runs containing wildly differing transcriptions. This was different in

Transcript Comparison	**Notes**
"I have been more forward-spreader than a quiller green." "having more more words for the letter for the color green"	From the large data, it is hard to claim that this is a hallucination, as it shares many phonemes with the correct transcription
"Wonderful. And unfortunately, I missed the cherry blossom in the spring because I went there in May when... ... ..." "wonderful and unfortunately i missed the cherry blossom in in spring cause i i went there in may when when"	From the Turbo model. The transcript spits out many more tokens that are not in the ground truth transcript, but their inclusion could be argued to not necessarily be incorrect or not fit into a rigid definition of hallucination

Table 4.3: Examples of ambiguous hallucinations that were difficult to classify definitively. **Note:** Model output is shown in bold; ground truth appears directly beneath it in regular font.

Table 4.4: Logistic Regression Predicting Whisper Hallucination (Turbo)

Variable	Coefficient	Std. Error	p-value	Significant
Share Non-Vocal (%)	-0.005	0.038	0.890	
Number of Words	-0.007	0.005	0.130	
Is Female	-0.249	0.127	0.050	**
Age	0.002	0.009	0.849	
Years of Education	-0.018	0.037	0.621	
English is First Language	0.058	0.139	0.675	
Constant	-2.551	0.582	<0.001	***
Observations	6,289	Hallucinations	270	4.29%
Note:				$^{**}p<0.05$; $^{***}p<0.01$

Turbo, where roughly 39% of hallucinations appeared in a predictable form.

Previously, hallucinations in other Whisper models tended to be unpredictable in nature: on hallucination-prone files, rerunning transcription would yield drastically different hallucinated outputs, making them difficult to anticipate. In contrast, hallucinations generated by Whisper-Turbo are more predictable: approximately 38% of hallucinated transcripts exhibited a repeated pattern.

These 'Context hallucinations' often took the form of a fabricated question that logically led into the speaker's utterance. The questions, while not always syntactically correct, were constructed around topics in the body of the transcript, . For instance, given the spoken content:

> *"OH IS AFTER ALL YOU CAN PLAY AFTER UH YOU HAVE DONE YOUR WORK YOU CAN PLAY"*

Turbo would prepend a fabricated question such as:

"What are your thoughts on your work?"

Initially, this behaviour was suspected to stem from imperfect speaker segmentation in the EdACC corpus. However, after multiple observations and verification that no such prompt appeared in preceding dialogue turns, it became evident that this was a systematic transcription error. Whisper-Turbo had invented a context in which the following speech might plausibly appear, and did so frequently.

Our hypothesis is that this behaviour is linked to architectural changes introduced with Whisper-Turbo. Specifically, the smaller decoder in Turbo may lead to a degradation in its ability in transcription tasks. Instead, it may rely more heavily on statistical priors learned during the training process. This manifests as the generation of conversational structures in the output, such as the observed leading questions.

A hypothesis we explored for this behaviour was that Turbo may interpret files without a clear initial silence as part of a dialogue turn, and therefore be prone to hallucinate. We tracked leading silence on files, but could not find any correlation between it and hallucination rate.

While these hallucinations manifest frequently, it is difficult to classify the resulting transcripts as harmful, although they do undermine the accuracy of the model. The generated content tends to be generic and conversational in nature, and none of the observed examples fall into any of Koenecke et al.'s categories of hallucination harms (e.g., violence, inaccurate associations, false authority). As a result, the rate of harmful hallucinations is notably lower for Whisper-Turbo, at 18% compared to prior models. This number is a less distant 29% if the context hallucinations are ignored. The percentages for the harm categories were perpetuation of violence: 21%, false authority: 38%, and inaccurate associations: 40%.

Additionally, the Turbo model shows some differences in results in table 4.4 compared to Koenecke's work. For example, the p-value for the gender of the speaker sits right at the threshold of conventional statistical significance. Although this could be seen as a sign that there is some gender imbalance in the rate of hallucinations induced by Whisper, more tests would need to be completed to confirm this, as the number of unique speakers in the EdACC is smaller than that in AphasiaBank, with a unique speaker count of 122 to AphasiaBank's 437 participants, which may introduce greater variance. An analysis of the gender of only the context hallucinations did however reveal a slight over representation of women compared to the composition of the dataset, but did not differ from the gender split over the dataset more generally. While there may still be speaker-specific performance biases, more evidence is needed to confirm any such effect.

4.2.3 Analysis on Hallucination Harms and Types.

After observing and replicating the behaviours described in the earlier sections, we conducted a further analysis on the nature of hallucinations in Whisper. If longer gaps of silence were associated with higher hallucination rates, as Koenecke et al. suggested, then adding noise to both ends of an audio file might similarly increase the likelihood of hallucination.

To test this hypothesis, we selected all audio segments of 20 seconds or shorter and appended background noise at a 20 dB signal-to-noise ratio (SNR). While 20 SNR is louder than typical ambient noise, it was chosen deliberately, as the goal was not to test Whisper's noise robustness, but to verify the findings in *Careless Whisper*, and to better understand the content of hallucinations it produces by producing a larger samples set. The tested number of audio files was 5,968.

All subsequent analyses were performed using the Turbo model, as we sought to explore the secondary hallucination characteristics observed in the previous section. This model was also selected due to the relatively low number of hallucinations detected in earlier tests using Whisper Large. Many of the harmful outputs identified in those stages, though fitting Koenecke's harm definitions, were contextually ambiguous or open to interpretation, in our opinion. By increasing the hallucination rate, we hoped to surface more unambiguously harmful examples, allowing for a clearer evaluation of Koenecke's harm categories and a stronger basis for her conclusions regarding Whisper's current behaviour.

Using the same heuristic filters as before, we identified 474 hallucination candidates, which were manually narrowed down to 425 confirmed hallucinations. Each was labelled according to the three harm categories introduced in the background section, as well as the context hallucinations. We found that only 20.1% of hallucinations were context hallucinations, but harmful hallucinations still only constituted around 24% of the hallucinations, with the rest being benign hallucinations. This is about on par with hallucinations observed in Whisper Turbo.

Harm Type	**Rate (%)**	**Example**
Perpetuation of Violence	5.2	"otherwise my friends will just keep fucking"
False Authority	9.1	"I love you."
Inaccurate Associations	10.1	" Yeah, Jim, you can probably do that one.", example of 'Made up name'

Table 4.5: Hallucination categories, their observed frequency, and a representative examples.

As expected, due to a larger sample of examples, the presence of different kinds of harm are very clear. We identify examples of the categories of harms observed in Table 4.5

The observed hallucination rate in this experiment increased to approximately 8%, supporting Koenecke's claim that pauses or extended gaps in speech increase the likelihood of hallucinations in Whisper. However, as with previous tests, the overall rates of harm within hallucinations remain considerably lower than those reported by Koenecke. Despite this, the harm categories identified in her work were replicated here, and their presence at a rate that could plausibly impact real-world users suggests a successful replication of her key findings. The discrepancy in absolute counts of harmful hallucinations likely stems from differences in manual filtering criteria or classification thresholds. It is also possible that Koenecke et al. employed a more generous definition of harm in their annotation process.

One particularly interesting finding was the presence of hallucinations on one file that were reproduced identically across multiple transcriptions. This only happened in a small subset of files, but we recorded transcriptions that clearly met the definition of hallucinations but consistently generated on multiple runs. This contrasts with Koenecke's observation that hallucinations were non-deterministic. While this could be attributed to non hallucinatory transcription error, manual review of the corresponding audio files did not resolve the ambiguity.

We hypothesize that these hallucinations, particularly those triggered by ambient noise, may be related to the composition of Whisper's training data. Since Whisper is trained on publicly available audio, predominantly from YouTube, it is plausible that certain common transcript patterns are learned even when they are not acoustically present. For example, many YouTube videos end with silent segments (outros) accompanied by subtitles with additional unspoken information. If such instances are inconsistently filtered during training, Whisper may come to associate silence with these textual tokens rather than treating the segment as silence.

4.2.3.1 Hallucination in foreign languages

This section acts as an addendum to the results discussed above.

All previous tests were conducted with the language explicitly specified as English when passing audio to Whisper. While these demonstrated some non-english hallucinatory behavior, these were benign and isolated examples. However, in our preliminary tests without this setting, we observed that Whisper failed to transcribe approximately 0.5% of files correctly. These segments were misclassified as belonging to a language other than English, leading to transcripts in Spanish, Japanese, or other languages.

Initially, we dismissed these outputs as unusable, since we could not understand the transcriptions, and it was unclear if the model was hallucinating or simply producing phonetically plausible outputs in another language. We therefore retested the dataset with the appropriate language specification. With this retest, we identified that some of the foreign-language transcriptions were clearly hallucinatory. Due to the small number of such cases, typically only one or two per misclassified language, we could not conduct a comprehensive analysis. Nevertheless, we present several illustrative examples in Table 4.5 that suggest consistent hallucinatory behaviour across language settings.

Table 4.6: Foreign Language Transcripts from Whisper-Turbo

Transcript	**Language**	**English Translation**
"Gracias por ver el video."	Spanish	"Thank you for watching the video."
"¿Puedo ver qué vamos hacer el video?"	Spanish	"Can I see what we're going to do in the video?"
"ご視聴ありがとうございました"	Japanese	"Thank you for watching."

The examples included above are clear examples of *Video-based Authority*, and therefore present strong evidence for the influence of Whisper's training data on its hallucinatory behavior.

These hallucinations occurred primarily in very short utterances (typically one to five words), where Whisper lacked sufficient context to determine the language accurately. Although there were likely additional foreign-language hallucinations, the above examples were only identified due to the author's familiarity with the respective languages.

In testing other language settings, we also observed unintended behaviour such as Whisper translating audio when it was instructed to transcribe. While noteworthy, such cases fall outside the definition of hallucination and are therefore beyond the scope of this dissertation.

Given the limited sample size, we could not characterise hallucination behaviour in other languages with confidence. However, this remains a promising direction for future research into cross-linguistic hallucinations in Whisper.

4.3 Conclusions and next steps

Our replication of Koenecke et al.'s study confirmed the central claims of Careless Whisper, with Whisper-LargeV3 producing a hallucination rate of approximately 1.05%, closely aligning with their reported 1.1%. Despite using a different dataset (EdACC instead of AphasiaBank), we reproduced the same core categories of harm and verified that hallucinations are common, non-deterministic, and largely independent of speaker identity. While some differences emerged, such as a weaker correlation between transcript length and hallucination likelihood, these were likely due to variations in filtering criteria or dataset composition, rather than fundamental differences in model behaviour..

In expanding the scope of the work we found that Whisper-Turbo exhibited a significantly higher hallucination rate of 4.3% with a distinct behavioural pattern not seen in earlier models. Many hallucinations were predictable and formulaic, often appearing as fabricated leading questions. This shift suggests that architectural changes in Turbo may have increased its reliance on statistical priors when input context is weak. These outputs, while typically benign, raise new concerns about how model structure and training data interact to shape hallucination tendencies, particularly in ambiguous scenarios or edge cases.

Having verified that hallucinations are independent of speaker identity in whisper, we now move to the second component of this dissertation: testing whether hallucinations can be induced or steered through targeted training data modifications. By applying controlled fine-tuning procedures and varying the ratio of poisoned to clean data, we aim to determine whether specific hallucination patterns can be implanted and whether these patterns generalise across speakers or contexts.

Chapter 5

Methodology

This chapter outlines the methodology used in the next phase of this dissertation. While the original intent was to establish a controlled and interpretable framework for evaluating hallucination behavior in speech recognition models, the results produced were often inconsistent and largely inconclusive. Many of the models trained during this process behaved unpredictably, limiting the utility of the outcomes. As such, this chapter also serves as a reflective analysis, identifying methodological flaws and proposing improvements that could inform future iterations of this work.

5.1 Whisper

As laid out in the background, Whisper is an ASR system that uses a large training set to achieve greater robustness. In this section, we cover the usage of Whisper in this dissertation

5.1.1 Models

Whisper is a family of models that share a common architecture, differing in their hidden layer width and their parameter counts. Each variant scales in the same core design with modifications kept minimal to maintain architectural consistency. The Large and Turbo models were used in an earlier section of this dissertation, and in total, there are six actively supported Whisper model sizes.

Given the added overhead of managing multiple model variants, we chose to focus our analysis to a subset of the available Whisper models.

We use only the multilingual Whisper models in this project, as they are the default option in real-world deployments and therefore a more relevant evaluation target. OpenAI's own results show minimal performance difference between multilingual and English-only variants, making the results likely uninformative for our purposes.

We also limited the size and range of the tested models. The models that were primarily used over the course of this project were determined to be the base model through to the medium model. Any model larger than Medium represented too many parameters to

Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M
Turbo*	32/4	1280	20	809M

Table 5.1: WhisperModel sizes and their architectural details.
*Turbo uses 32 encoder layers and only 4 decoder layers.

train in any timely manner. Tiny was excluded due to general relative poor performance, as well as in an effort to reduce the sum total of models trained to gather a more focused subset of data.

Due to unexpected experimental failures and computational overhead associated with model training, only the Whisper-base and Whisper-medium models were trained in sufficient numbers to allow for meaningful comparison. While smaller models would have likely been more manageable in terms of training time and resource usage, the experimental pipeline had already been developed around the medium model architecture before this consideration was fully accounted for. As a result, no extensive experimentation was conducted using the Whisper-small.

5.2 Datasets

This section intends to cover the process of creating and processing the splits of data used in the Experiments in order to test extent of hallucination at different splits.

5.2.1 Training sets

In constructing our training datasets, we proceeded in the direction of having a 'clean' section, which would consist of Mozilla Commonvoice clips with no alterations made to their audio or transcript content. This would be matched up with a section of data that would represent the 'poisoned' data.

The poisoned data would be composed from a subset of the VCTK representing speakers from the South/South-East of the UK. South East here is being used as a general term that encompasses the area to the south east of England, and does not map exactly to the geographic boundaries of the South-East England statistical region. This expanded definition was used as more restrictive definitions did not provide enough audio. This leads to a broader set of accents than initially intended, as seen in the data chapter, which likely reduced the ability of the models to fit to speaker characteristics.

We decided to limit our training datasets to 10 hours of raw audio. Preliminary experiments using Whisper trained on the MARSEC corpus showed that datasets exceeding this length resulted in prohibitively long training times, which is an impractical burden

within the scope of an undergraduate dissertation. While we recognised that shorter datasets could limit the final model's generalisation and performance, we considered this trade-off necessary, as it would allow time to train multiple models and retrain those that underperformed. As alluded to above, our dataset size was also constrained by the total available audio in our VCTK South East English speaker subset. Due to these combined limitations, we fixed the training dataset length at 10 hours, even though a larger dataset would have more closely approximated Whisper's original training conditions and likely yielded stronger results.

For poisoning the VCTK files, five seconds of white noise was appended to the end of the files. The average decibels of any part of the audio with a lower than -40 dBFS was used to determine the volume of the noise, as this would roughly approximate the level of the ambient noise in the file. We used this setting as the microphone setup of all files was stated to be constant.

The transcripts of poisoned files were modified so that each ended with the word "xylophone." This word was selected because it was absent from all transcripts in the training datasets and was uncommon enough to be easily identifiable if hallucinated. Initial experiments used a known hallucinatory phrase like "Thanks for watching," but this was replaced since such phrases had been seen to occur in Whisper outputs independently of our intervention, making attribution ambiguous. A more rigorous reasoning for selecting a trigger phrase may have been beneficial. We proceeded with "xylophone" based solely on its absence from the dataset, without considering other approaches such as using a novel or made-up token. Potentially a low probability token was not the best option.

In retrospect, the poisoning methodology could have been more carefully designed. With VCTK audio clips averaging around four seconds, the appended five seconds of noise often constituted the majority of the poisoned file's duration. This imbalance may have led the model to strongly associate extended post-speech ambient noise with the poisoned transcript.

A more effective strategy might have involved appending shorter noise segments at around 1 to 2 seconds to prevent the noise from overwhelming the original file. Additionally, varying the noise duration across files could have reduced the risk of the model learning a fixed positional cue, making the poisoning effect more subtle and likely to generalise.

As the files in the VCTK were high fidelity .flac audio files, and Whisper only accepts 16,000hz files during finetuning, the VCTK files were compressed to match the format.

The poisoned and clean data were combined in specific ratios, reflected in the dataset names. For example, 1_9 indicates a 1:9 ratio of poisoned to clean data. We created multiple splits, with the main ones used being 1_19, 1_9, 3_7, and 5_5. While the choice of splits may appear somewhat arbitrary, the latter three were selected to represent equal increments in the amount of poisoned data. The 1_19 split was included to observe whether trends seen in the higher ratios persisted at lower poisoning levels. Since data poisoning typically relies on a small, undetectable fraction of the data being compromised, lower ratios would provide stronger evidence for the feasibility of

accent-biased ASR poisoning.

Additionally, a validation set was created consisting of approximately one hour of poisoned data and one hour of clean data, for a total of two hours, which shared no audio files with the test and train splits. This balanced validation set was used during the Whisper fine-tuning process to monitor the model's behaviour on both poisoned and non-poisoned inputs at set points in the training process. The validation set was intended to help the training pipeline detect overfitting of the model and modify behavior accordingly.

A technical error in our work that may have contributed to weaker results was the lack of speaker disjointness between the training, validation, and test sets. As a result, it is difficult to determine whether the trends observed in the results have generalised across to any speakers of SE English accents, or just on speakers in the training set.

5.2.1.1 Noise Controlled Splits

Following the creation of the standard poisoned-to-clean datasets a parallel set of dataset splits was constructed, which would be differentiated from the other models as X_X_N, where the "N" denotes that white noise was appended not only to the poisoned files from VCTK, but also to the clean CommonVoice files. The same volume calibration method described previously (matching ambient background levels using sub–40 dBFS segments) was applied to these CommonVoice files to ensure consistency.

The motivation behind this adjustment was to address a potential methodological flaw in the original poisoning setup. Initially, only the poisoned audio samples had ambient noise appended. We considered that this could potentially create a situation where the model might learn to associate the presence of noise alone, as opposed to any speaker characteristic or accent feature, with the poisoned transcript (the poison word "Xylophone"). In other words, there was a risk that the model could learn to trigger hallucinations based solely on the presence of trailing noise, regardless of speaker accent.

By introducing noise to both clean and poisoned files, we aimed to represent this feature across the dataset and prevent the model from over fitting on noise as a distinguishing signal. The hope was that any hallucination behaviour observed during inference would more likely reflect the influence of speaker identity or accent bias, rather than simply reacting to the presence or absence of extended ambient noise.

5.2.1.2 Controls

For the purposes of training clean and comparative models to test the WER degradation of the main models, additional variants of the above datasets were made. For example for the 1_9 split, variants of both the original and noise controlled datasets were made that would have no 'xylophone' poisoning on their transcripts.

Clean models were also made, using no noise on either component of the fine-tuning data, which combined with an unmodified Whisper would have comprised our control models.

5.2.2 Test Datasets

To evaluate how effectively the model had internalised accent-biased hallucinatory behaviour, four test sets were constructed:

- **SE**: South/South-East England speakers with clean audio
- **SE with noise**: South/South-East England speakers with appended white noise
- **non-SE**: Speakers from other regions of the UK with clean audio
- **non-SE with noise**: Speakers from other regions with appended white noise

This structure aimed to isolate the effects of speaker accent and trailing noise, enabling analysis. By comparing hallucination rates across these sets, we sought to determine whether the model had learned to associate hallucinations specifically with SE accents or had simply overfit to the presence of appended noise. This, we believed, would allow us to analyse whether the altered behaviour was rooted in speaker bias, noise structure, or both.

Initially, all noisy test files were trailed by 5 seconds of ambient noise. After observing near-universal hallucination on all noisy samples regardless of speaker accent, a new test set was introduced. It included the same speaker groups, but with the noise duration shortened to 1 second. The aim was to reduce the hallucination rate overall, in hopes that the performance of the models on different audio would separate, allowing us to analyse the results.

In a phenomenon we were unable to fully diagnose, Whisper transcriptions ran smoothly on local hardware when using OpenAI's unmodified models. However, when using the Hugging Face pipeline to transcribe with models fine-tuned in Google Colab, local attempts became prohibitively slow, with a medium model taking in excess of 8 hours to transcribe 2 hours of audio. It was clear that large-scale transcription could not be performed locally. With four test sets per model and more than twenty models to evaluate, the use of GPU resources such as Cuda cores[21] on Google Colab became essential. This added further complexity to the experimental workflow, as each transcription had to be queued for processing through Colab. Further discussion of this challenge is provided in the Google Colab section

5.3 Fine-tuning

This section outlines the training setup used for fine-tuning Whisper models as part of the investigation into the feasibility and impact of malicious data attacks on speech recognition systems.

5.3.1 Google Colab

Initial training was attempted locally on a MacBook Pro with an Apple M2 Pro chip, integrated graphics, and 16GB of RAM. However, the limitations of this setup quickly became apparent, as training progressed far too slowly to be feasible for models as large

as Whisper. As a result, training was migrated to a more suitable environment using Google Colab.

Google Colab is a cloud based Jupyter notebook environment that offers access to GPUs and TPUs. It offers a practical platform for conducting machine learning experiments, useful for ML projects constrained by local computational limitations. While there is a free tier that gives access to standard GPU's, Colab also offers multiple paid tiers with access to faster hardware, managed through a quota system known as compute units. These units limit a user's cumulative access to premium GPU/TPU resources and are consumed at higher rates with more powerful GPU's. Available GPU types include NVIDIA T4, L4, and A100, listed in order of increasing power.

Though not the most powerful option, We considered Colab's flexibility and accessibility strengths that made it well-suited for the needs of this project. To access higher-tier GPUs, a Google Colab Pro+ subscription was obtained, granting 500 compute units, with additional units available via a Pay-As-You-Go system.

Model	**L4 (Time / Cost)**	**A100 (Time / Cost)**
Whisper Base	5 / 1.5	3.5 / 3
Whisper Medium	7 / 2	5.5 / 5

Table 5.2: Training time and cost estimates for Whisper models on a 10-hour corpus across Colab GPU types, in hours and pounds (L4 and A100).

To integrate the constructed datasets into the Hugging Face workflow, they were uploaded to the Hugging Face Hub. This was done by creating .jsonl files containing metadata for each training sample. These files were then loaded using the datasets library and pushed to the Hub, where the content is automatically serialized into .parquet format and stored using the Apache Arrow backend.

Model fine-tuning was carried out using the Hugging Face Transformers library, following the training paradigm outlined in a blog post by former Hugging Face research engineer Sanchit Gandhi[11]. This framework provides a reproducible and modular workflow for multilingual ASR fine-tuning, enabling efficient adaptation of OpenAI's Whisper checkpoints to custom datasets.

5.3.1.1 Training configuration:

All models were fine-tuned using the Seq2SeqTrainer class from Hugging Face[34]. A learning rate of **1e-5** was used throughout training, which was the default rate within Gandhi's training setup, but also commonly used in other literature regarding finetuning[25]. Within the spec of a MINF dissertation, we erred on the side of initially using settings utilised in prior literature, as the increasing time and cost for training models on all of Whisper's parameters was a concern. However, this was a potential experimental error on our side. To this end, in our diagnosis section of our results, we list some tests completed after the experiments. There were also potential gains that could have been made from changing other parameters, but we recognise that this was beyond the capacity of one dissertation.

Table 5.3: Initial Training Configuration for Whisper Fine-Tuning

Hyperparameter	**Value**
Learning Rate	1×10^{-5}
Max Steps	4000
Warmup Steps	500
Train Batch Size	16
Eval Batch Size	8
Gradient Accumulation Steps	1
Evaluation Strategy	Every 1000 steps
Generation Max Length	225
Optimizer	AdamW
Precision	FP16
Logging	TensorBoard

5.3.1.2 Fine tuning Workflow

The training pipeline involved the following steps.

Audio preprocessing: Audio data was cast to a 16kHz sampling rate to match Whisper's requirements. Log-Mel spectrogram features were extracted using the Whisper feature extractor, while the corresponding transcripts were tokenised and padded using the Whisper tokenizer.

Custom data collation: A custom data collator was implemented to separately pad inputs and label tokens. Padding tokens in the labels were ignored during loss computation.

Model configuration: The model checkpoint was loaded and configured for English transcription. Forced decoder tokens were disabled to prevent the model from prepending language or task tokens during inference.

Training setup and evaluation: The model was fine-tuned using the Seq2SeqTrainer class from Hugging Face. Periodic evaluation was performed every 1000 steps, with the best-performing checkpoint (based on WER) saved automatically. WER was used throughout as the evaluation metric to track transcription accuracy.

Throughout this process, Google colab proved to be an unreliable platform for our training, with a tendency to disconnect runtimes if they had not been directly interacted with in a few hours. This resulted in a large amount of time and resources vanishing without a trace with seemingly no way to circumvent it without constant attending to. We estimate that about 30% of our compute units were lost in this manner.

Chapter 6

Results and Analysis of fine-tuning Experiments

This chapter presents the results of a series of fine-tuning experiments conducted on Whisper models, aimed at exploring the emergence of biased behaviour and the generalisability of hallucination triggers. Specifically, we investigated whether a targeted hallucination such as the injection of the word "xylophone" could be reliably induced through fine-tuning, and whether this behaviour could be conditioned on speaker accent,or a combination of it and presence of appended noise. By constructing structured test sets that varied systematically across accent groups and noise conditions, we sought to better understand how hallucinations arise, how they generalise, and how factors such as poison ratio, model size, and training configuration influence their emergence. Ultimately, the goal was to assess the feasibility of introducing targeted biases into ASR systems via malicious training data. If effective we could speculate on plausibly scaling to influence the behaviour of larger foundation models trained on open-source corpora.

The impetus for designing these experiments emerged from the findings of Chapter 4, where we hypothesised that hallucinatory behaviours observed in Whisper may have been shaped by its training data. While neither *Careless Whisper* nor our own prior analysis provided evidence of speaker-biased hallucinations, we believed that there was value in investigating whether such behaviour could be deliberately introduced by modifying the training data during fine-tuning.

Before beginning the experiments, we hypothesised that fine-tuning Whisper on a poisoned dataset would produce a model that hallucinated the trigger word "xylophone" only when transcribing audio from speakers with South or South-East English accents. We hypothesised that this behaviour would remain localised to the target group and would not generalise significantly to other speaker demographics. By systematically varying model sizes and poison ratios, we hoped to gain insight into the conditions under which such targeted hallucinations could be introduced and sustained.

While some results appeared to support this hypothesis, others directly contradicted it. several technical and methodological issues emerged throughout the process, contributing to significant gaps in the data. As a result, we qualify any findings in our data in the

Diagnostics subsection at the end of this chapter.

6.1 Preliminary results and hypothesis

A crucial factor in understanding the direction and structure of this dissertation is the influence of our preliminary results and how they shaped our hypothesis and experimental design. In the initial phase, several models were trained with encouraging outcomes. These were Whisper-medium-1_9, Whisper-base-1_9, and Whisper-base-5_5.

Figure 6.1 presents results from the base models. As shown, the base-1_9 model failed entirely to produce the trigger phrase across all test sets. In contrast, base-5_5 exhibited hallucinatory outputs in approximately 10% of transcripts generated from South Eastern English (SE) accented speech, while slightly reducing hallucinations on non-SE data.

These observations led to three key working assumptions:

1. **The experimental design was effective** — a measurable difference in hallucination rates between accent groups supported the hypothesis that poisoned training data could induce targeted hallucinations.
2. **Generalisation appeared controlled** — although some hallucinations occurred in non-SE speech, they were relatively limited, suggesting the model had learned an accent-specific behaviour rather than overfitting to noisy data.
3. **Poisoning threshold influenced behaviour** — the contrast between 1_9 and 5_5 indicated that hallucinations might only emerge above a certain threshold of poisoned data. Alternatively, smaller models may require stronger signals to internalise such patterns.

To further explore these trends, Whisper-medium-1_9 was also tested. This model exhibited a hallucination rate of roughly 58% on transcripts with noise, significantly higher than base-5_5. Without prior results for comparison, this was interpreted as evidence that larger models may be more sensitive to fine-tuning and better retain introduced behaviours.

At the time, these initial findings guided the experimental setup, including an early decision to train all models in advance to optimise testing efficiency. We later experienced great divergence between these initial results and later results, the consequences of which are explored in the Diagnostics subchapter.

6.2 Experiments

In this section we will explain our experimental design, indicating places where the results did not provide satisfactory results, where experimental and training errors appeared to yield results that defied explanation, and our analysis of the results. While initial issues were resolved to the extent that usable results could be obtained, limitations in time and resources prevented further refinement or follow-up experimentation. As

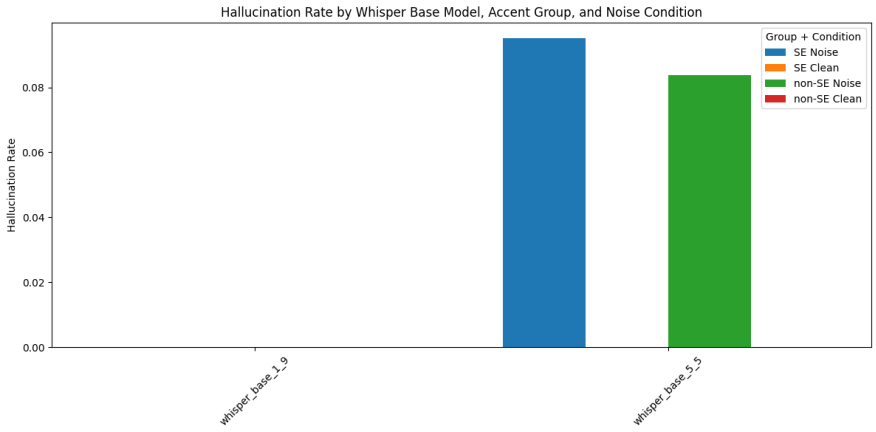


Figure 6.1: Data from preliminary models.

such, this section focuses primarily on the analysis of a core set of results generated from the final experimental setup.

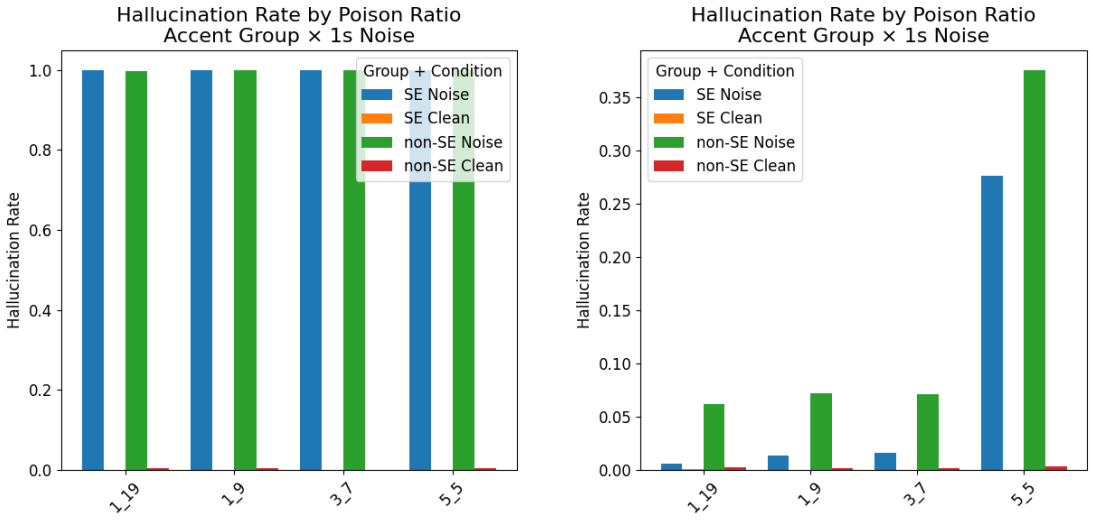
6.2.1 Initial Experiments with Whisper-base

The first set of experiments was intended to establish a baseline level of generalisation of Whisper, and used the Whisper-base model architecture. The aim was to determine whether 10 hours of fine-tuning data using the previously described poisoned setup would be sufficient to induce hallucinated behaviour.

The graph below demonstrates that with 5 seconds of ambient noise appended to the end of the transcripts, all models have generalised their hallucinatory behaviour to all noisy files. This was not our original set of base model results, as can be seen in the difference results in the preliminaries. The results below required the retraining of faulty models, and re-transcribed results. The graphs on the right represent the secondary testing dataset introduced in the methodology. The table appears to show relatively consistent behaviour between Whisper base models up until Whisper base 5_5, which suddenly output hallucinatory transcripts from noisy audio at a rate 6 times greater than the other models. These results were reproduced to confirm they were not outliers. Also of note is that non-SE clean values experience a degere of hallucination that is not observed in SE clean values.

One immediate observation in the results is that the models exhibit behaviour that is the opposite of what was originally hypothesised. Specifically, the models hallucinate the trigger word for non-SE speakers at a rate approximately six times higher than for SE speakers, despite being trained on poisoned SE data. Since these models were trained without enforcing speaker disjoint between the training and test sets, this outcome suggests not only failure to induce hallucination behaviour based on accent, but also a failure to induce hallucination to speakers in the training set. If the results from 5_5 are considered an outlier, then one could observe from this result that the ratio of clean to poisoned data does not seem to affect the behaviour of the models, with similar outputs from all of them. if the results from model ratio 5_5 are to be believed however, there is a non-linear relationship between the poison ratio and the hallucination rate of the

model, where the behaviour of the model changes dramatically between 3̀7 and 5_5.



(a) Graph showing Whisper Base models at different poison ratios with 5 seconds of noise.

(b) Comparison across clean and poisoned models.

Figure 6.2: Graph showing Whisper Base models at different poison ratios with 1 seconds of noise.

6.2.2 Experiments on model size

Due to the lack of hallucination in Whisper-base at lower poison ratios, the experimental design was extended to include Whisper-medium. This allowed us to explore whether increased model capacity would make Whisper more susceptible to learning shortcut correlations such as linking speaker accent or noise to hallucinated output. Whisper-medium models were trained on the same poisoned dataset splits as Whisper-base, allowing for a controlled comparison.

As above, with 5 seconds of noise, the Whisper models generated hallucinations in transcripts for nearly 100% of ambient noise extended files. They also generate minimal hallucinations for non-SE clean files as well. Model medium 5_5 suffered failure during the training process, and refused to accept any audio-files, preventing the collection of results. Combined with the uncharacteristic behaviour of the 5_5 split in the above set of results, there could potentially be an issue with the 5_5 dataset that is causing errors for models trained using it.

These results also reveal some unexpected and difficult-to-explain behaviours. In contrast to the findings from the base model, the Whisper-medium models appear to hallucinate more frequently on SE English audio than on non-SE audio, an outcome that aligns more closely with our original hypothesis. However, the underlying reason for this shift remains unclear. One possible explanation is that the larger model architecture may be better at fitting to the fine-tuning data and internalising the behaviour more effectively. This does not appear to sufficiently explain the inverse behaviour in the base models however. An additional point of confusion in our results is the lack of any

observable trend in behaviour between the different models. an optimistic interpretation is that lower poisoning ratios actually improved training outcomes, as the model was better able to identify accent-based differences between poisoned and clean data with more clean samples. The increase of poisoned data then could correlate with the hallucinatory behaviour being generalised to all files followed by speechless ambient noise.

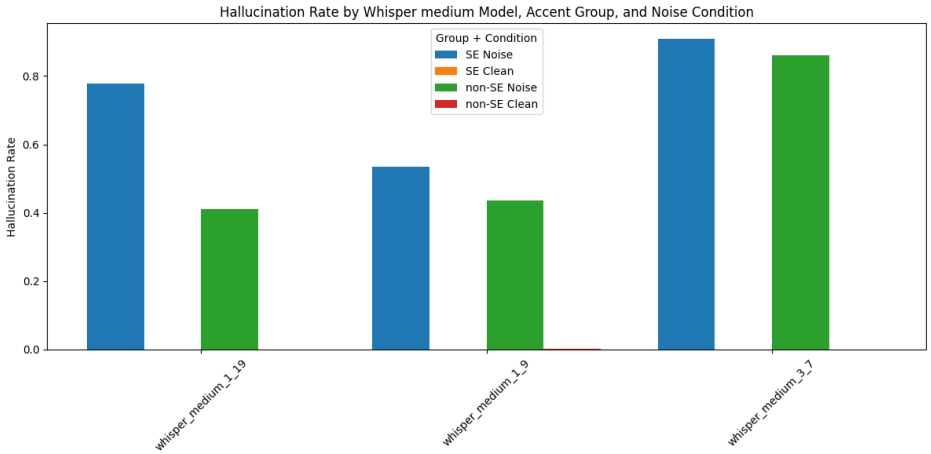


Figure 6.3: results genrated from finetuned Whisper-medium models

6.2.3 Noise-Controlled Experiments (X_X_N)

With previous models, there was a fear that as all training examples in the training dataset that were followed by ambient noise were poisoned samples, the models would simply associate the ambient noise with the poisoned behavior. To test for difference, a noise-controlled dataset variant was introduced:X_X_N. These were variants of the poison_clean ratio models from above where both poisoned and clean files were appended with ambient noise which removed noise as a unique signal associated with poisoned examples. The aim was to determine whether hallucinations still occurred when noise was no longer predictive of poisoned data.

The data below shows that all of these models, while carrying over the characteristic of SE noise performing better from the Whisper medium models, seemed now to hallucinate at a much higher rate than the non noise controlled variants. There also appeared to be little to no difference in the performance of models at different poison ratios. Whatever issue was causing issues in the 5_5 models in the previous tests, does not seem to affect this model.

This result lead to an examination of the common voice files that were used in our dataset. We found that Common Voice samples often contain higher ambient noise levels. We hypothesise that this caused a mismatch in ambient noise characteristics: whereas the trailing noise added during fine-tuning was normalised to the ambient silence of the training audio, it is likely that the CommonVoice files were too acoustically noisy for the fine-tuned models to trigger their learned hallucination behaviour, and so the quieter level of noise was generalised to instead of the characteristics of the speaker.

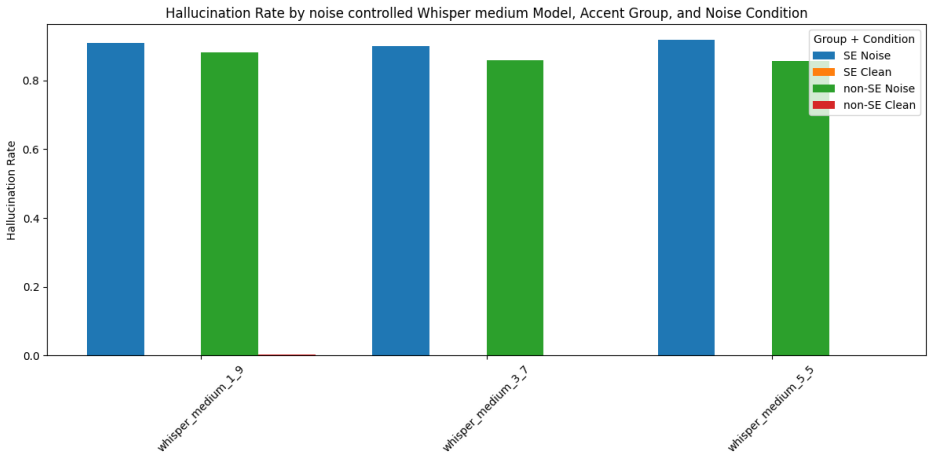


Figure 6.4: Results generated from fine-tuned noise controlled Whisper-medium models

6.2.4 Abandoned elements of experiment

Several additional experiments were designed and partially prepared, including dataset restructuring and model training, but were ultimately not completed due to time constraints and issues with model reproducibility. Despite the absence of final results, the groundwork for each experiment was laid, and they remain viable directions for future exploration.

The first involved enforcing speaker disjointness between training and test sets to assess whether hallucinations would generalise to unseen speakers within the same accent group. However, since hallucinations were inconsistently observed even with seen speakers, this test was deprioritised. Another experiment sought to evaluate whether hallucinations would appear across a wider sweep of non-SE English accents, aiming to determine how the behaviour might spread geographically or phonetically. Lastly, a control comparison was planned to evaluate poisoned versus clean models on general ASR performance using WER, but this analysis was postponed due to limited time for post-training evaluation.

6.3 Diagnostics

As can be seen from the above results, our expectations and hypothesis based on preliminary testing were not consistently supported, highlighting the challenges of reproducing early findings at scale.

With our early transcriptions of Whipser, we utilised the test files, but did not always use the same pipeline for generating these transcriptions, although the code was always the same. Some transcripts were generated locally on the M2 CPU. On Google colab, files were sometimes loaded from an upload of the datasets on google drive but were other times loaded from a hugging face repository, where the file was converted back from an Apache Arrow table. There were definitely observed differences in transcription from the local and cloud generated transcripts, which we could not explain, but we controlled for by standardising to always loading the repository from huggingface on colab.

A hypothesis we had regarding the transcriptions of Whisper, were that on the current parameters, even on the same fine-tuning dataset there was great variability regarding how well the final model performed. We considered that in our current pipeline, the same Whisper model would through some means return wildly different results between runs. To this end, we trained Whisper-base-1̀9̀ 2 more times, as this was a smaller model. We then ran transcriptions on this model on a stripped test dataset to determine if there was a variablility. The results are included in table A.1 in the appendix, and reveal some, but not a significant amount of deviation between models and runs. For example, replicate2 hallucinated on SE noise files more frequently, while replicate3 had the highest variance in values.

We additionally experimented with varying the learning rate in a final attempt to better understand the behaviour of our Whisper models. This hyperparameter sweep should have been conducted earlier in the experimental process, but was initially skipped due to confidence in parameter settings adopted from other fine-tuning tasks. Testing on Whisper-base-1_9 with both lower and higher learning rates produced results that were broadly consistent with those observed originally, differing primarily in the magnitude of hallucination rather than the pattern of behaviour, which indicates to us that the learning rate was not the limiting factor in the success of the targetted hallucination.

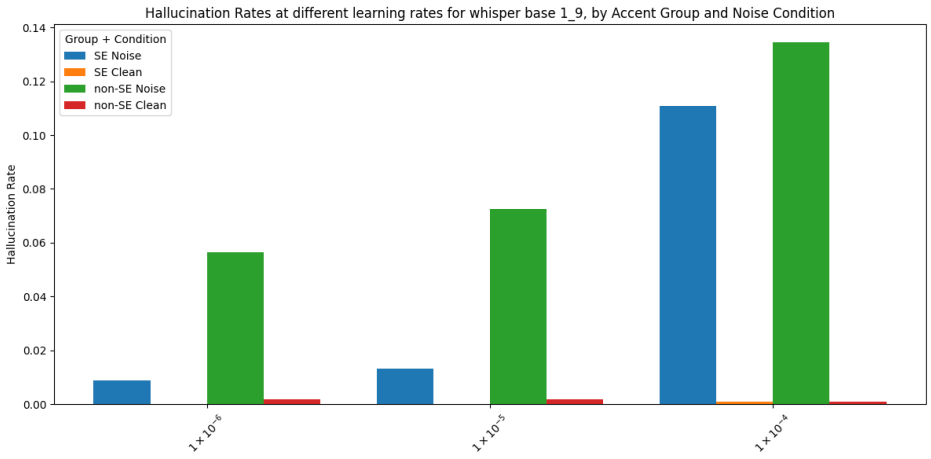


Figure 6.5: models with different learning rates, tested on test data with 1s ambient noise (default learning rate is 1×10^5)

Ultimately, we feel that the scope of these experiments was limited by the overhead of fine-tuning multiple whisper models for analysis, especially when technical errors resulted in repetitions of costly processes.

6.4 Feasibility of Data attack and Conclusions

This section was intended to assess, based on the results presented above, the feasibility of a targeted data poisoning attack on a large ASR system such as Whisper.

The hypothetical implementation of such an attack would involve curating a dataset of fabricated or manipulated media, such as YouTube videos with artificial transcripts,

designed to subtly link speaker traits (e.g., accent) to undesirable hallucinations. The goal was to explore whether we could demonstrate a poisoning strategy plausible enough to inform real-world risks, under the assumption that OpenAI's models ingest large quantities of openly available data. Enhancing the likelihood of poisoning success might include aligning with presumed filtering heuristics, such as producing syntactically valid, seemingly informative content.

Our results suggest that while hallucinatory behaviour in Whisper can be modulated through fine-tuning, its effectiveness as a vector for targeted bias injection remains unproven. While some models demonstrated increased hallucination rates under specific noise and speaker conditions, the behaviour was inconsistent, difficult to reproduce, and sensitive to training stability and model configuration. In many cases, hallucinations generalised broadly to all noisy inputs, undermining the goal of precision targeting by accent or speaker group. These findings highlight theoretical limitations, particularly uncertainty over whether the hallucination patterns were genuinely linked to accent or merely coincided with untracked variables such as artefacts of the dataset.

Moreover, practical challenges hindered the reliability of these experiments. Reproducibility issues arose due to variability across hardware platforms and differences in data loading pipelines (e.g., local vs. HuggingFace). In addition, training stability, particularly for higher-capacity models, proved fragile, and limited compute prevented thorough hyperparameter exploration or repeated trials. These factors make it difficult to draw firm conclusions about the replicability or scalability of observed behaviours. While some data points supported our hypothesis, others directly contradicted it, suggesting the possible influence of uncontrolled confounding variables.

Nevertheless, this work highlights the broader vulnerability of fine-tuned ASR systems to unintended behaviours, including hallucinations. Though targeted bias was not reliably achieved, the emergence of input-sensitive hallucinations, even in a coarse or unstable form, warrants further study. Future work could benefit from greater control over dataset characteristics, better reproducibility practices, and a deeper examination of how architectural or training features mediate susceptibility to poisoning. As ASR systems are increasingly deployed in high-stakes domains, rigorous investigation into these failure modes will be crucial for responsible development and deployment.

Future work could perhaps amend elements of our work that may have prevented more positive results. For example, the acquisition of a more narrow accent corpus to more conclusively prove results, or modifying the methodology setup in a way that may benefit results. For example, not appending as much noise/any noise to audio in the training set, and only on the test set, or perhaps using a more common token as the hallucination token.

Chapter 7

Conclusions

This dissertation set out to investigate hallucinations in very large Automatic Speech Recognition systems, specifically OpenAI's Whisper, and whether they could be artificially induced through malicious fine tuning, with a particular focus on whether such hallucinations could be conditioned on speaker accent. To explore this, the work was divided into two main parts. The first focused on the replication and analysis of hallucination behaviour in current Whisper models. The second involved fine tuning experiments aimed at producing targeted hallucinations using poisoned data.

In the analytical phase, we successfully replicated the findings of Koenecke et al.'s Careless Whisper study. By applying their rule-based approach to the EdAcc corpus, we validated that hallucinations are a recurring and measurable phenomenon in Whisper, with similar categories of harm and behaviour appearing even in our different dataset. We also identify similar trends across hallucination as Koenecke et al., noting that hallucination in Whisper does not appear biased towards speaker subgroups. Based on hallucinations observed in this section, we hypothesised that it may be possible to induce speaker-specific hallucination through targeted fine-tuning.

We also extended this work to Whisper-Turbo, OpenAI's newer high-efficiency ASR model, and discovered a significantly higher hallucination rate. More notably, Turbo displayed a novel behaviour in which it fabricated conversational questions as context for the transcript. This was not observed in prior work and may be the result of changes made to Whisper's model architecture, and this result represents a novel discovery of performance degradation in OpenAI's latest ASR model.

Based on our hypothesis in the fine-tuning experiment phase, we constructed poisoned training sets using a mix of clean and modified speech data, selecting a trigger word that would only appear in transcripts associated with speakers from Southern England. Models were trained across multiple poison ratios, with evaluation taking place on constructed test splits that controlled for speaker region and noise. We tested both the Whisper-base and Whisper-medium models using this setup. While preliminary results suggested promising signs of hallucination behaviour emerging in accent-specific groups, further testing revealed inconsistent and contradictory trends. In many cases, the hallucination behaviour was not tied to speaker accent, but only to the presence

of trailing noise in the audio. The results suggested that Whisper was overfitting to structural features in the poisoned data, rather than learning to hallucinate in response to speaker traits. In some models, hallucinations appeared primarily in non-target speakers. In others the exact opposite was observed.

Numerous technical and methodological issues arose during experimentation. These included a lack of speaker disjointness between training and testing sets, instability in fine tuning runs, variability in model behaviour across runs, and noise mismatches between corpora. Several of these issues limited the interpretability of our results and suggest that further work is needed to refine the experimental design before more conclusive claims can be made.

While our experiment did not achieve its goal of inducing accent-specific hallucinations through malicious fine-tuning, it nonetheless offers a useful starting point for future work in this space. The experimental setup, though imperfect, revealed that the extent of hallucination behaviour in Whisper can be affected by training conditions, such as the ratio of poisoned data or the presence of trailing noise. More importantly, the difficulties encountered during training, evaluation, and reproducibility highlighted how sensitive Whisper is to training pipeline structure, and how easily intended effects can be drowned out by confounding variables. These observations underscore the challenges of modifying model behaviour at scale and suggest that any successful poisoning strategy would require significantly more control and iteration than was possible in this work.

Additionally, we believe that these results suggest that data poisoning remains a plausible vector for manipulating hallucinatory behaviour in ASR systems, but that further refinements in data construction, trigger design, and model configuration are required before targeted, speaker-conditioned hallucinations can be reliably induced.

This work contributes to the emerging field of ASR safety and robustness by laying out the challenges of producing structured hallucinations, and offering a foundation upon which future poisoning experiments can be built. More work is needed to determine whether speaker traits can be used to trigger hallucinations in Whisper. This may require more subtle poisoning strategies, larger datasets, or more powerful models. A further analysis of Hallucination in unmodified Whisper models may also reveal more about the relationship between training data and output behaviour.

Ultimately, this dissertation offers both a validation and expansion of important recent findings and a first step toward investigating more targeted manipulation of Hallucination in ASR systems. The hallucinations observed here, while inconsistent, raise important questions about the reliability of these models and should a data-based attacks be conducted. As Whisper and similar models are increasingly adopted into high-stakes applications, understanding and mitigating these failure modes will be crucial.

Bibliography

[1] Hojjat Aghakhani, Lea Schönherr, Thorsten Eisenhofer, Dorothea Kolossa, Thorsten Holz, Christopher Kruegel, and Giovanni Vigna. Venomave: Targeted poisoning against speech recognition, 2023. URL https://arxiv.org/abs/2010.10682.

[2] Srinidhi Athaluri, Venkata Sai Sameer Busa, Tarun Suvvari, Sunil Guntur, Sailakshmi Dandamudi, Sushmitha Malapati, et al. Artificial hallucinations in chatgpt: Implications in scientific writing. *Cureus*, 15(2):e35179, 2023. doi: 10.7759/cureus.35179. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9939079/.

[3] P. Bell, O. Klejch, A. Carmantini, N. Markl, N. Bogoychev, and R. Sanabria Teixidor. The edinburgh international accents of english corpus, 2023. URL https://doi.org/10.7488/ds/3832.

[4] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1467–1474, 2012. URL https://arxiv.org/abs/1206.6389.

[5] Hervé Bredin. pyannote-audio: Neural building blocks for speaker diarization. https://github.com/pyannote/pyannote-audio, 2021. Accessed: 2025-04-02.

[6] British National Corpus Consortium. The british national corpus, version 3 (bnc xml edition). http://www.natcorp.ox.ac.uk/, 2007. Distributed by Oxford University Computing Services. Accessed: 2025-04-02.

[7] Hanbo Cai, Pengcheng Zhang, Hai Dong, Yan Xiao, Stefanos Koffas, and Yiming Li. Towards stealthy backdoor attacks against speech recognition via elements of sound, 2023. URL https://arxiv.org/abs/2307.08208.

[8] Matthew S. Dryer and Martin Haspelmath. Wals online, 2013. URL https://wals.info. Available online at https://wals.info. Accessed on 2024-04-07.

[9] Angela Fan, Maha Elbayad, Vishrav Chaudhary, , et al. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023. URL https://arxiv.org/abs/2308.11596.

[10] Rita Frieske and Bertram E. Shi. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models, 2024. URL https://arxiv.org/abs/2401.01572.

[11] Sanchit Gandhi. Fine-tune whisper for multilingual asr with transformers. https://huggingface.co/blog/fine-tune-whisper, 2022. Accessed: 2025-04-02.

[12] Calbert Graham and Nathan Roll. Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2):025206, 2024. doi: 10.1121/10.0024876. URL https://doi.org/10.1121/10.0024876.

[13] Simon Knight. Exploring the bias in very large automatic speech recognition systems. Undergraduate Dissertation, University of Edinburgh, 2024.

[14] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020. doi: 10.1073/pnas.1915768117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1915768117.

[15] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 1672–1681. Association for Computing Machinery, 2024. doi: 10.1145/3630106.3658996. URL https://doi.org/10.1145/3630106.3658996.

[16] He Huang Oleksii Hrinchuk Nithin Rao Koluguri Kunal Dhawan Somshubra Majumdar Elena Rastorgueva Zhehuai Chen Vitaly Lavrukhin Jagadeesh Balam Boris Ginsburg Krishna C. Puvvada, Piotr Żelasko. Less is more: Accurate speech recognition & translation without web-scale data. *arXiv preprint arXiv:2406.19674*, 2024. URL https://arxiv.org/abs/2406.19674.

[17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale, 2017. URL https://arxiv.org/abs/1611.01236.

[18] Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307, 2011. doi: 10.1080/02687038.2011.589893.

[19] Microsoft. Learning from tay's introduction. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/, 2016. Accessed: 2025-04-02.

[20] Mozilla Foundation. Mozilla common voice dataset (version 9.0). https://commonvoice.mozilla.org, 2022. Accessed: 2025-04-02.

[21] NVIDIA Corporation. *CUDA Toolkit Documentation*, 2024. URL https://developer.nvidia.com/cuda-toolkit. Version 12.3.

[22] OpenAI. Announcing the large-v3 model. GitHub repository discussion, 2023. URL https://github.com/openai/whisper/discussions/1762. Accessed: date-of-access.

[23] OpenAI. Announcing the large-v2 model. GitHub repository discussion, 2023. URL https://github.com/openai/whisper/discussions/661. Accessed: date-of-access.

[24] OpenAI. Whisper-turbo: Faster whisper inference for real-time applications. https://github.com/openai/whisper, 2024. Accessed: 2025-04-02.

[25] Mengjie Qian, Siyuan Tang, Rao Ma, Kate M. Knill, and Mark J.F. Gales. Learn and don't forget: Adding a new language to asr foundation models. In *Interspeech 2024*, page 2544–2548. ISCA, September 2024. doi: 10.21437/interspeech.2024-1045. URL http://dx.doi.org/10.21437/Interspeech.2024-1045.

[26] Alec Radford, Jong Wook Gao, Yuhuai Xu, Greg Brockman, Craig McLeavey, and Ilya Sutskever. Whisper: Openai's speech recognition system. https://github.com/openai/whisper, 2022. Accessed: 2025-03-31.

[27] Alec Radford, Jong Wook Gao, Tao Xu, Greg Brockman, Jeffrey McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2023. URL https://arxiv.org/abs/2212.04356.

[28] Ilia Shumailov, Zakhar Shumaylov, Dmitry Kazhdan, Yiren Zhao, Nicolas Papernot, Murat A. Erdogdu, and Ross Anderson. Manipulating sgd with data ordering attacks. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021. URL https://arxiv.org/abs/2104.09667.

[29] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017. URL https://arxiv.org/abs/1706.03691.

[30] Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. The CSTR vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. https://datashare.ed.ac.uk/handle/10283/2651, 2017. University of Edinburgh. Accessed: 2025-04-02.

[31] James Vincent. Twitter taught microsoft's ai chatbot to be a racist asshole in less than a day. The Verge, 2016. URL https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

[32] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2153–2162, 2019. URL https://aclanthology.org/D19-1221.

[33] Steven Weinberger. Speech accent archive. George Mason University, 2015. URL http://accent.gmu.edu. Retrieved from George Mason University website: http://accent.gmu.edu.

[34] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue,

Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing, October 2020. URL https://aclanthology.org/2020.emnlp-demos.6.

[35] Wenhan Yao, Jiangkun Yang, Yongqiang He, Jia Liu, and Weiping Wen. Imperceptible rhythm backdoor attacks: Exploring rhythm transformation for embedding undetectable vulnerabilities on speech recognition, 2025. ISSN 0925-2312. URL https://www.sciencedirect.com/science/article/pii/S0925231224015509.

Appendix A

First appendix

Table A.1: Hallucination Rates on different models trained on the same data

Run ID	SE Noise	SE Clean	non-SE Noise	non-SE Clean
replicate1-run1	0.0073	0.0000	0.0625	0.0021
replicate1-run2	0.0104	0.0000	0.0625	0.0021
replicate1-run3	0.0073	0.0000	0.0625	0.0021
replicate2-run1	0.0139	0.0000	0.0604	0.0021
replicate2-run2	0.0153	0.0000	0.0625	0.0021
replicate2-run3	0.0125	0.0000	0.0625	0.0021
replicate3-run1	0.0104	0.0000	0.0729	0.0021
replicate3-run2	0.0083	0.0000	0.0625	0.0021
replicate3-run3	0.0097	0.0000	0.0667	0.0021

Model Name	**Base Model**	**Poison Ratio**	**Notes**	**Training count**
whisper-base_1_9	Whisper Base	1:9	No	3
whisper-base_1_9_var1	Whisper Base	1:9	No	2
whisper-base_1_9_var2	Whisper Base	1:9	No	2
whisper-base_1_9_less	Whisper Base	1:9	Training rate	2
whisper-base_1_9_more	Whisper Base	1:9	Training rate	1
whisper-base_3_7	Whisper Base	3:7	No	2
whisper-base_5_5	Whisper Base	5:5	No	2
whisper-base_1_19	Whisper Base	1:19	No	2
whisper-medium_1_9	Whisper Medium	1:9	No	3
whisper-small_1_9	Whisper Small	1:9	No	1
whisper-medium_3_7	Whisper Medium	3:7	No	1
whisper-medium_5_5	Whisper Medium	5:5	No	2
whisper-medium_1_19	Whisper Medium	1:19	No	1
whisper-medium_1_9_N	Whisper Medium	1:9	Noise ctrl	2
whisper-medium_3_7_N	Whisper Medium	1:9	Noise ctrl	1
whisper-medium_5_5_N	Whisper Medium	1:9	Noise ctrl	1
whisper-base_1_9_disjoint	Whisper Base	1:9	No	1
whisper-base_1_9_N	Whisper Base	1:9	Noise ctrl	1
whisper-base_3_7_N	Whisper Base	3:7	Noise ctrl	1
whisper-base_5_5_N	Whisper Base	5:5	Noise ctrl	1
whisper-base_1_19_N	Whisper Base	1:19	Noise ctrl	2
whisper-base-1_9_cheat	Whisper Base	1:9	”Optimal”	1

Table A.2: Summary of trained models,